(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **06243013 A**

(43) Date of publication of application: **02.09.94**

(51) Int. Cl **G06F 12/00**
G06F 13/00

(21) Application number: **05054788**

(22) Date of filing: **19.02.93**

(71) Applicant: **TOSHIBA CORP**

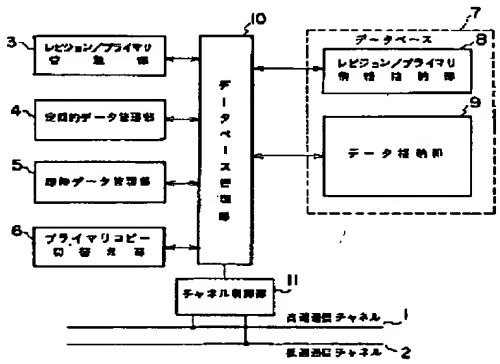
(72) Inventor: **SAKURA OSAMU**

(54) DISTRIBUTED DATA BASE SYSTEM

(57) Abstract:

PURPOSE: To eliminate the need of excessive communication for a synchronous processing, to prevent an update stop at the time of a site fault and communication network interruption and to prevent the deterioration of the processing by dividing a primary copy into appropriate size and dynamically dispersing the places of the divided copies between sites.

CONSTITUTION: A revision/primary management part 3 executes revision management and primary management. When its own site is not the primary copy and the present revision of the page differs from the latest revision of the page in the whole data base system, the announcement of effect that the page of its own site to be the primary copy is transmitted to the whole system by a high speed channel. At that time. the site where the page is the primary copy compares the revision in the middle of the announcement with the latest revision and transmits data of the difference by using the high speed channel. Data of the page of its own site is updated by the data.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平6－243013

(43)公開日　平成６年(1994)９月２日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　12/00	５３３　Ｊ	8944－5B		
13/00	３５５	7368－5B		

審査請求　未請求　請求項の数２　ＦＤ　（全 10 頁）

(21)出願番号　特願平5－54788

(22)出願日　平成５年(1993)２月19日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　櫻　修
東京都府中市東芝町１番地　株式会社東芝府中工場内

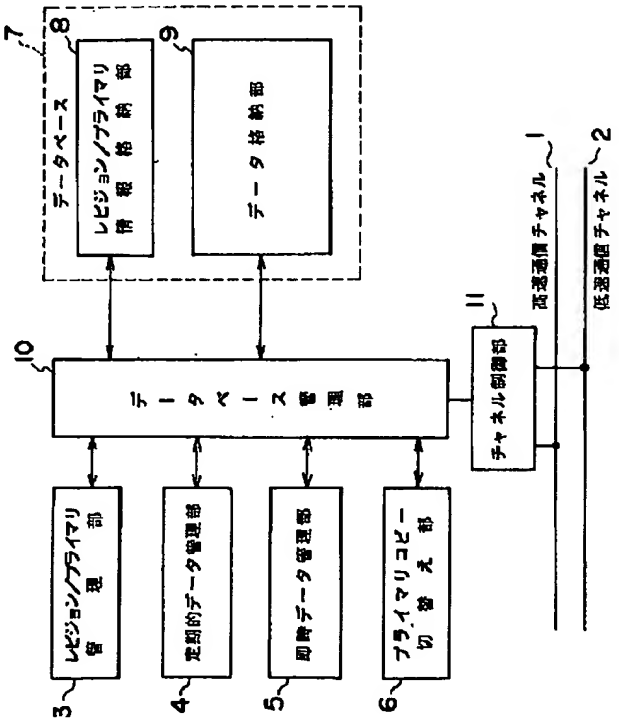
(74)代理人　弁理士　石井　紀男

(54)【発明の名称】　分散型データベースシステム

(57)【要約】

【目的】　分散型データベースシステムにおいて、同期処理のための余分な通信を要せず、サイト障害時及び通信網分断時の更新停止がなく、更に処理効率の低下をなくす。

【構成】　複数のデータベース間を結ぶ高速，低速の各通信手段と、ページ毎のレビジョン管理及びプライマリ管理を行なう手段と、定期的なデータの等化を行なう手段と、アクセス時に最新データがなかった場合の即時データ等化を行なう手段と、書き込み／更新時にプライマリコピー切替えを行なう手段とよりなり、読み込み時、自サイトの当該ページのレビジョンが最新レビジョンより古ければ、高速通信チャネルにより最新データをプライマリサイトへ要求し、書き込み／更新時、自サイトの当該ページのレビジョンが最新レビジョンより古ければ、高速通信チャネルにより最新データをプライマリサイトへ要求し、最新データ受信を同時に自サイトの当該ページをプライマリコピーとして全サイトへ宣言し、又、アクセスとは無関係に全サイトにつき定期的にデータの等化を行なう。



1

【特許請求の範囲】

【請求項１】　複数のデータベース間を結ぶ高速，低速の各通信手段と、ページ毎のレビジョン（Revision）管理及びプライマリ管理を行なう手段と、定期的なデータの等化を行なう手段と、アクセス時に最新データがなかった場合の即時データ等化を行なう手段と、書き込み／更新時にプライマリコピー切替えを行なう手段とよりなり、読み込み時、自サイトの当該ページのレビジョンが最新レビジョンより古ければ、高速通信チャネルにより最新データをプライマリサイトへ要求し、書き込み／更新時、自サイトの当該ページのレビジョンが最新レビジョンより古ければ、高速通信チャネルにより最新データをプライマリサイトへ要求し、最新データ受信と同時に自サイトの当該ページをプライマリコピーとして全サイトへ宣言し、又、アクセスとは無関係に全サイトにつき定期的にデータの等化を行なうことを特徴とする分散型データベースシステム。

【請求項２】　プライマリコピーの切替えはページ毎のアクセス回数が所定値に達したときのみ行なう手段を備えたことを特徴とする請求項１記載の分散型データベースシステム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、分散型データベースシステムに関し、特に各データベースサイトに設けた複数データベースの資源管理に関する。

【０００２】

【従来の技術】図６は分散型データシステムを説明する全体システム図であり、図の例ではＡ，Ｂ，Ｃからなる３つのデータベースサイトが通信網501 で接続され、分散型データベースを構築している。各データベースサイトはサーバー502 ，データベース７，端末装置503 よりなる。分散型データベースシステムでは、複数のサイトが同一データを重複して持つことが可能である。分散型データベースシステムの様々なサイトで特定のデータの検索がある場合、レスポンスタイムの改善、通信費の削減のため、同じデータを重複して複数のサイトに格納したいという要求がある。又、信頼性を向上させるため、複数のデータベースサイト間で同一のデータを重複して持ち、データベースサイトの障害に備えたいという要求もある。しかし、従来の技術では複数のデータベースサイト間の同期処理，等化処理のため、かえって通信量（費）が増えたり、処理効率（レスポンス）が低下したり、あるいは障害時に更新が停止したり、といった問題が逆に生じてしまう。

【０００３】

【発明が解決しようとする課題】以下、これら従来技術による問題点を説明する。前述したように、同じデータを重複して複数のサイトに格納した場合、データが検索される場合はよいが、更新される場合の複数のデータベースサイト間で通信が発生することになる。即ち、複数のデータベースサイトに重複して格納されているデータの内容を一致させることを要し、お互いに更新処理を同期させる余分な通信が必要になる。又、分散型データベースシステムでのデータベースの重複は、システムの信頼性向上に効果があるものの、処理の効率の観点からは余分な通信が発生し、処理効率が低下する欠点がある。読み出し専用のデータベースシステムであれば重複は問題とならない。しかし、書き込みがある場合には、各トランザクションは任意のコピーに対して操作を行なうことになるため、同時実行制御などトランザクション管理が重複データをもたない場合に比べて複雑になる。重複データにおける同時実行制御としては、従来から種々の方法があるが、ここでは本発明の基礎ともなっているプライマリコピー（元帳）方式について説明する。

【０００４】図７はプライマリコピー方式を説明する図であり、この方式は複数の重複データベースのなかで、特定の１つをプライマリコピーとして特別に扱う。読み／書きを行なうためにはこのコピーをロックしておけば、全てのコピーに対してロックする必要がなくなる。図７において、例えばトランザクションＢが書き込み／更新を行なう場合、まず、プライマリコピーのロック状態を調べ、もしロックされていなければ、これをロックして書き込み／更新をする（引き続き、書き込み／更新された内容が各サイトのコピーに反映されることとなるが、これについてはデータベース管理システム（ＤＢＭＳ）の機能に依存する。）。又、プライマリコピーがロック中であれば、トランザクションＢはロックが解除されるまで待つ。又、読み出しの場合には自サイトのコピーから読み取る。この方式は（書き込み／更新頻度）＜（読み出し頻度）の場合に有効である。

【０００５】重複データを持つ分散型データベースシステムでは、更新についても問題が複雑である。全部のコピーを同時更新する方法では、サイトに障害があった場合、全く更新不能になってしまう。この対策として、「更新できるデータベースサイトを更新し、更新の履歴を保存する。障害を起こしたデータベースサイトは、処理を再開する前にこの履歴に従って、ほかのコピーとの同期をとる。」という方法があるが、プライマリコピーを持つサイトが障害を起こした場合にはこの方法では対策とならない。他の問題として、通信網の一部が故障して、互いに通信できない分離したサブ通信網へ分割されてしまう場合がある。図８がこの場合を示している。この場合、個々のサブ通信網内では、通信可能であるため処理を続けることができる。しかし、夫々のサブ通信網内で独立にコピーを更新すると矛盾が生じる。この問題に対する１つの解決法としても、プライマリコピー方式を用いることができる。即ち、プライマリコピーを持つサブ通信網だけがそのデータの更新を行なうことができ、ほかのサブ通信網はそのデータの古い値を読むこと

ができる。本発明は上記事情に鑑みてなされたものであり、同期処理のための余分な通信を要せず、サイト障害時及び通信網分断時の更新停止がなく、更に処理効率の低下のない分散型データベースシステムを提供することを目的としている。

【０００６】

【課題を解決するための手段】本発明の［請求項１］に係る分散型データベースシステムを図１によって説明する。なお、図１は１つのデータベースサイトの構成図である。図１において、１は高速通信チャネル、２は低速通信チャネルであり、これにより他のデータベースサイトと連結されている。３はページ毎のレビジョン管理及びプライマリ管理（自サイト名とプライマリサイト名がイコールならプライマリコピー）を行なう機能、４は定期的なデータの等化（更新履歴の伝送：低速通信チャネル使用）を行なう機構、５はデータベースアクセス時、最新データが無かった場合の即時データ等化（更新履歴の伝送：高速通信チャネル使用）を行なう機構、６は書き込み／更新時、プライマリコピー切替え（全サイトへの宣言：高速通信チャネル使用）を行なう機構である。又、７はデータベース、８はレビジョン情報及びプライマリ情報を格納しておく格納部、９はデータそのものの格納部、10はデータベース管理部、11はチャネル制御部である。本発明の［請求項２］に係る分散型データベースシステムは図５に示されており、ページアクセス計数部12を付加したことが前記［請求項１］との差異である。

【０００７】

【作用】本発明の［請求項１］に係る分散型データベースシステムは重複データの同期を行なうための方式として、従来技術のプライマリコピーにロックする方式を基礎とするが、従来のようにプライマリコピーを１つのサイトに固定して持つのではなく、プライマリコピーを適当なサイズ（ページと呼ぶ管理上の単位）に分割し、その所在をサイト間で動的に分散するようにしている。そして、書き込み／更新が生じた場合、そのサイトの持つページのコピーを（最新状態に更新した後）プライマリコピーとして全システムへ宣言する。先ず、レビジョン／プライマリ情報のフォーマットを図２によって説明すると、この情報（ページ毎）をもとにして、レビジョン／プライマリ管理部３がレビジョン管理及びプライマリ管理を行なう。レビジョン／プライマリ情報は、ページ＃部21，自サイトレビジョン部22，最新レビジョン部23及びプライマリサイト指示部24よりなる。各ページにはそのページの現在のレビジョン22と、データベースシステム全体でのそのページの最新レビジョン23と、そのページが現在プライマリコピーであるか否かを表す情報（及びプライマリコピーの所在サイト情報）24を持つ。データベースシステムの各サイトは、高速，低速の２つの通信手段（チャネル）１，２で結ばれている。システム全体での最新レビジョン及びプライマリコピーの切替え情報は、この高速チャネルによりシステム全体に伝達される。又、最新のデータそのものは低速チャネルにより伝送される（高速チャネルを使う例外の場合もある。後述。）。

【０００８】さて、読み／書き／更新を行なうときに、まず自サイトのページが現在プライマリコピーとされているか調べる。

ケース（ａ）

自サイトがプライマリコピーであれば、そのままで読み／書き／更新を行なって良い。但し、書き／更新を行なったときは、書き／更新後のそのページの最新レビジョンを高速チャネルでシステム全体に伝達する（このとき、書き／更新データそのものはすぐには伝送しない。低速チャネルを使って定期的に伝送される。又、他サイトからのプライマリ切替え要求があったときには高速チャネルを使って伝送される。）。

ケース（ｂ）

自サイトがもしプライマリコピーでなければ、そのページの現在のレビジョンと、データベースシステム全体でのそのページの最新レビジョンとを比較する。一致しており、かつ書き／更新の場合（読みはそのまま）、自サイトのそのページをプライマリコピーとする旨の宣言（一種のロック：ロック方式は何でも良い）を高速チャネルでシステム全体に伝達する。宣言が完了した後は、そのままで書き／更新を行なう。書き／更新後の処置は最初からプライマリコピーであった場合、ケース（ａ）と同じである。

ケース（ｃ）

自サイトがもしプライマリコピーでなく、かつそのページの現在のレビジョンと、データベースシステム全体でのそのページの最新レビジョンとが異なっていた場合も、自サイトのそのページをプライマリコピーとする旨の宣言を高速チャネルでシステム全体に伝達する（書き／更新の場合（読みはデータを送ってもらうのみ））。但し、この場合には現在そのページがプライマリコピーであるサイトは、宣言中の（これからプライマリになろうとしているサイトの該当の）レビジョンと最新レビジョンとを比較し、その差分のデータを高速チャネルを使って伝送してやる。このデータにより自サイトのページのデータが更新されてからの動作は、ケース（ｂ）と同じである。データはこのようにプライマリ切替え要求があったときだけではなく、定期的に低速チャネルを使って、又は夜間などの通信費の安価な時間帯に全サイトが最新状況となるように伝送される。

【０００９】その頻度は下記の例のような考えで最適化される。

目標平均レスポンス＞ケース（ｃ）発生時のデータ伝送量（平均）÷高速チャネルの伝送速度

目標最大レスポンス＞１ページのデータ量÷高速チャネ

ルの伝送速度（最悪ケースでは１ページ分のデータの伝送が最大値となる）

総通信費用＝（ケース（ｃ）の発生確率×ケース（ｃ）発生時のデータ伝送量×高速チャネル使用料）＋（定期データの伝送頻度×定期データ伝送時の伝送量×低速チャネル使用料（又は夜間などの特別料金））

（注：発生確率と伝送頻度の単位は同一）

ここで、ケース（ｃ）の発生確率及びケース（ｃ）発生時のデータ伝送量は、定期データの伝送頻度が多ければ多いほど小さくなる。しかし、ある頻度より多くすると、低速チャネル全体の使用料が高速チャネルのそれを上回ることになる。従って、通信費のみの観点からは、総通信費用が最低となる点（低速チャネル全体の使用料が高速チャネルのそれと同じになる点）が最適な点である。これにレスポンスにおける要求を加味して定期データの伝送頻度を決定する。又、特にデータベースシステムの信頼性向上の要求が高い場合には、ケース（ｂ）の発生確率が高くなるように定期データの伝送頻度を決定する。しかし、本発明では書き込み／更新が生じたサイトにプライマリコピーが移動することとなり、結果として自動的に最もそのページのデータに書き込み／更新を必要とするサイトにプライマリコピーが再配置される傾向があるため、伝送頻度はかなり低くても良い（従って、経済的である）。本発明の［請求項２］に係る分散型データベースシステムは、プライマリコピーの切替えは、ページアクセス計数部がカウントするページ毎のアクセス回数が所定の回数に達したときにのみに行なう。その他の作用は［請求項１］と同様である。

【００１０】

【実施例】以下図面を参照して実施例を説明する。図３は本発明の実施例の動作を説明するための図であり、各データベースサイトは図１で示した構成要素よりなる。そして各サイトにあるサーバー502 は、前述のページ毎のレビジョン／プライマリ管理部３，定期的なデータの等化部４，即時データ等化部５，プライマリコピー切替え部６，データベース管理部10及びチャネル制御部11よりなる。ここでは動作を主として説明するために各機構の作用線を別の記号で表す。即ち、

ａ．高速通信チャンネル。

ｂ．低速通信チャンネル。

ｃ．ページ毎のレビジョン管理及びプライマリ管理（自サイト名とプライマリサイト名がイコールならプライマリコピー）を示す作用線。但し、図３では引出線の端部に符号ｃを示すが、データベース内にあるものである。

ｄ．定期的なデータの等化（更新履歴の伝送：低速通信チャネル使用）を示す作用線。

ｅ．アクセス時、最新データがなかった場合の即時データ等化（更新履歴の伝送：高速通信チャネル使用）を示す作用線。

ｆ．書き込み／更新時のプライマリコピー切替え（全サイトへの宣言：高速通信チャネル使用）を示す作用線。

である。又、図４は図３のレビジョン／プライマリ情報の遷移を示す図であり、Ａ，Ｂ，Ｃの各サイトにあるものを夫々関連付けして並列的に示している。

【００１１】次に図３，図４を用いて作用を説明する。

今、サイトＡにおいて、あるページのデータを読み込もうとしている場合について説明する。この例では、ページ毎のレビジョン／プライマリ管理部の、自サイトレビジョン（値９０）と最新レビジョン（値９１）を比較すれば、自サイトのレビジョンが最新でないことがわかる。そこで、高速通信チャネルａにより、プライマリサイトＢに対して最新データを要求する。即ち、アクセス時、最新データが無かった場合は作用線ｅにて示されるように、即時データ等化（更新履歴の伝送）を行なう。プライマリサイトがＢであることはプライマリサイト指示部の情報（値Ｂ）によりわかる（元々最新ならば、そのまま読む）。

次にサイトＣにおいて、同じページの書き込み／更新をしようとしている。この例では自サイトのそのページのレビジョン（値８９）がやはり最新レビジョン（値９１）より古い。そこで、高速通信チャネルａによりプライマリサイトＢに対して最新データを要求する。即ち、前記同様に、アクセス時、最新データが無かった場合は、作用線ｅに示されるように即時データ等化（更新履歴の伝送）を行なう。データ受信と同時に自サイトのそのページをプライマリコピーとして全サイトへ宣言ｆする。これにより各サイトのこのページのプライマリサイト指示部は値○印となる（元々最新ならば、プライマリコピーの宣言のみ。元々プライマリならば、そのまま書き込み／更新）。

又、アクセスとは無関係に全サイトにつき定期的にデータの等化（更新履歴の伝送ｄ）がなされる（低速通信チャネルｂ使用）。これにより、全サイトのレビジョン／プライマリ情報は一致する（値９２，９２，Ｃ）。この定期的なデータ等化の頻度については、既に述べた通りである。

【００１２】上記実施例からわかるように、書き込み／更新が必要なサイトにプライマリコピーが自動的に再配置される。従って、運用の態様に応じて最適なプライマリコピーの配置が得られる。この結果、同期のための通信費の削減と、重複データベースであるにも拘らず、いつでもプライマリコピーが手元にあるようなレスポンスの実現が可能となる。又、そのサイトでのページ毎の書き込み／更新頻度に応じて当該サイトにプライマリコピーが集まる傾向となるため、障害時，通信網分断時についても、必要なページのプライマリコピーが手元にある確率が高く、運用に際して支障を生じにくい。

【００１３】図５は他の実施例の構成図であり、図１と同一部分については同一符号を付して説明を省略する。図１との相違点はページアクセス計数部12を設けたこと

である。そして、プライマリコピー切替え部６は、書き込み／更新の都度プライマリを切替えるのではなく、ページアクセス計数部12がカウントするページ毎のアクセス回数が、予め定めた一定の回数に達したときにのみプライマリを切替える。これにより、一時的なアクセスにより最適な配置が崩されるのを防ぐことができる。本実施例はたまたま更新されたというケースでの効率低下の防止となる。

【００１４】

【発明の効果】以上説明したように、本発明によれば自サイトで書き込み／更新の多いページのプライマリコピーは自動的に自サイトへ再配置される。即ち、書き込み／更新が必要なサイトにプライマリコピーが自動的に再配置され、運用の態様に応じて最適なプライマリコピーの配置が得られる。従って、重複データベースであるにも拘らず、どの利用者からみても、あたかもいつでもプライマリコピーが手元にあるようなレスポンス（処理効率）を実現すると共に、以下に列挙するような効果が得られる。

他サイトへのアクセスの減少。

ページのデータに書き込み／更新を必要とするサイトにプライマリコピーが再配置される傾向があるため、更新を停止しなければならない確率は低く、運用に支障を生じにくい。

自動的に最もそのページのデータに書き込み／更新を必要とするサイトにプライマリコピーが再配置される傾向があるため、通信網分断時にも必要なページのプライマリコピーが手元にある確率が高く、更新を停止しなければならない確率は低い。即ち、運用に支障を生じにくい。

【図面の簡単な説明】

【図１】本発明の原理ブロック図。

【図２】レビジョン／プライマリ情報フォーマット。

【図３】実施例の全体構成と動作を示す図。

【図４】図３の実施例のための、レビジョン／プライマリ情報の遷移を示す図。

【図５】他の実施例のブロック図。

【図６】分散型データシステムを説明する全体システム図。

【図７】プライマリコピー方式を説明する図。

【図８】通信網の分断を説明する図。

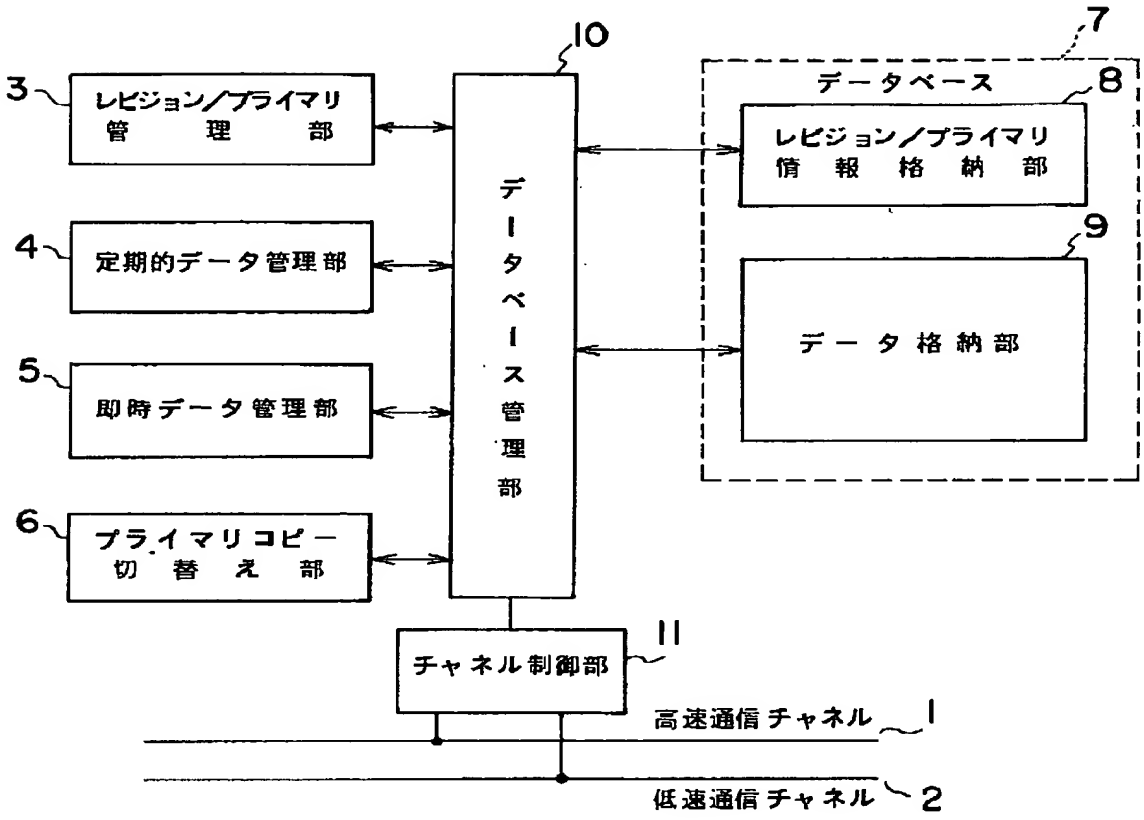
【符号の説明】

1　高速通信チャネル
2　低速通信チャネル
3　レビジョン／プライマリ管理部
4　定期的データ管理部
5　即時データ管理部
6　プライマリコピー切替え部
7　データベース
8　レビジョン／プライマリ情報格納部
9　データ格納部
10　データベース管理部
11　チャネル制御部

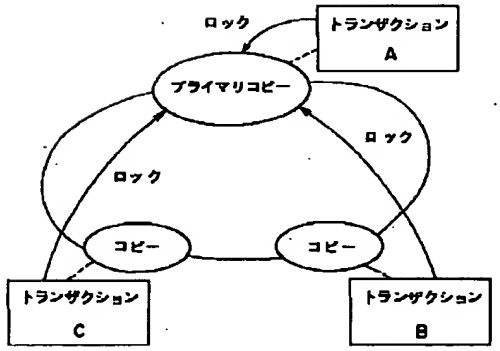
【図２】

ページ# (21)	自サイトレビジョン (22)	最新レビジョン (23)	プライマリサイト指示 (24)
００１			
００２			
００３			
⋮			

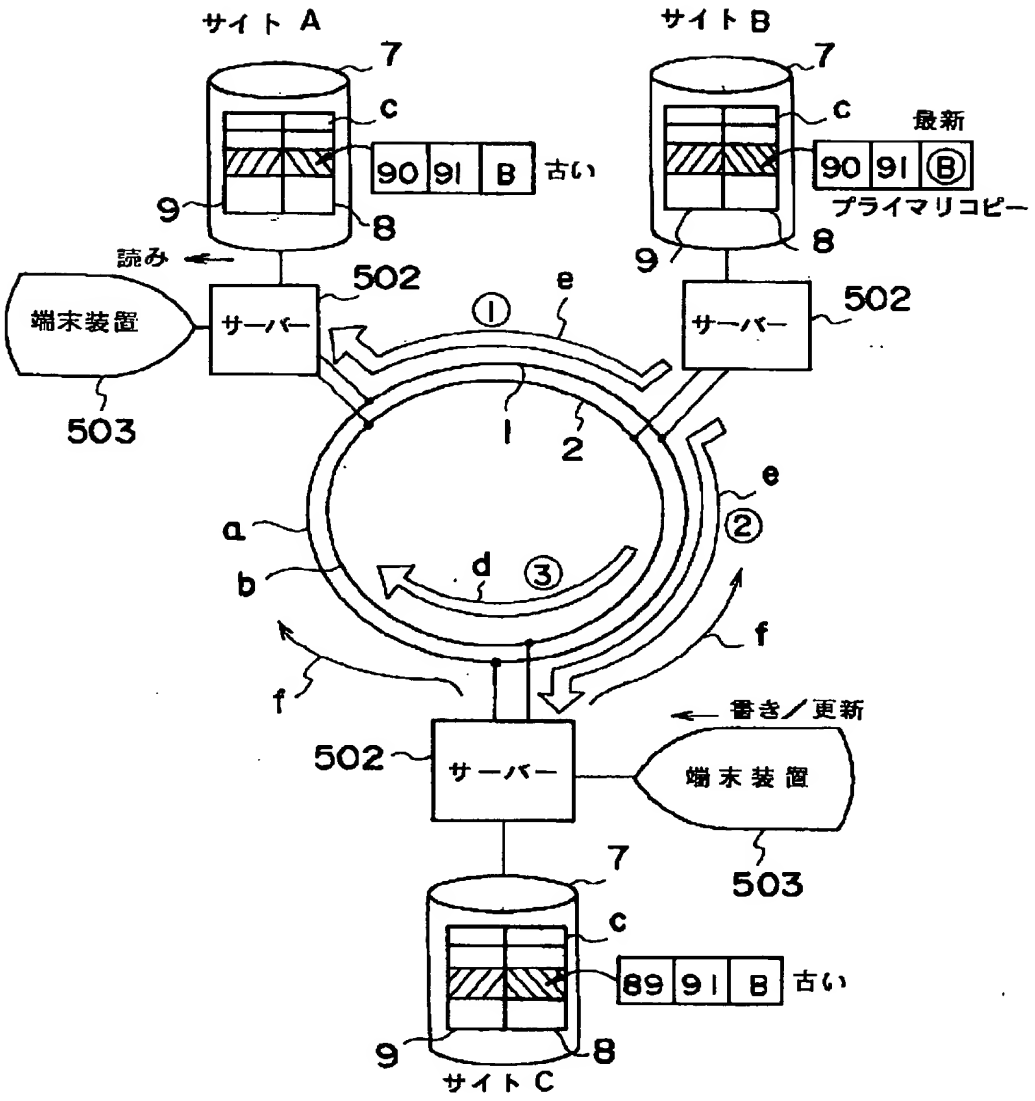
【図１】


10
7
データベース
8
3
レビジョン／プライマリ
管　理　部
レビジョン／プライマリ
情　報　格　納　部
4
定期的データ管理部
9
データベース管理部
データ格納部
5
即時データ管理部
6
プライマリコピー
切　替　え　部
11
チャネル制御部
高速通信チャネル
1
低速通信チャネル
2

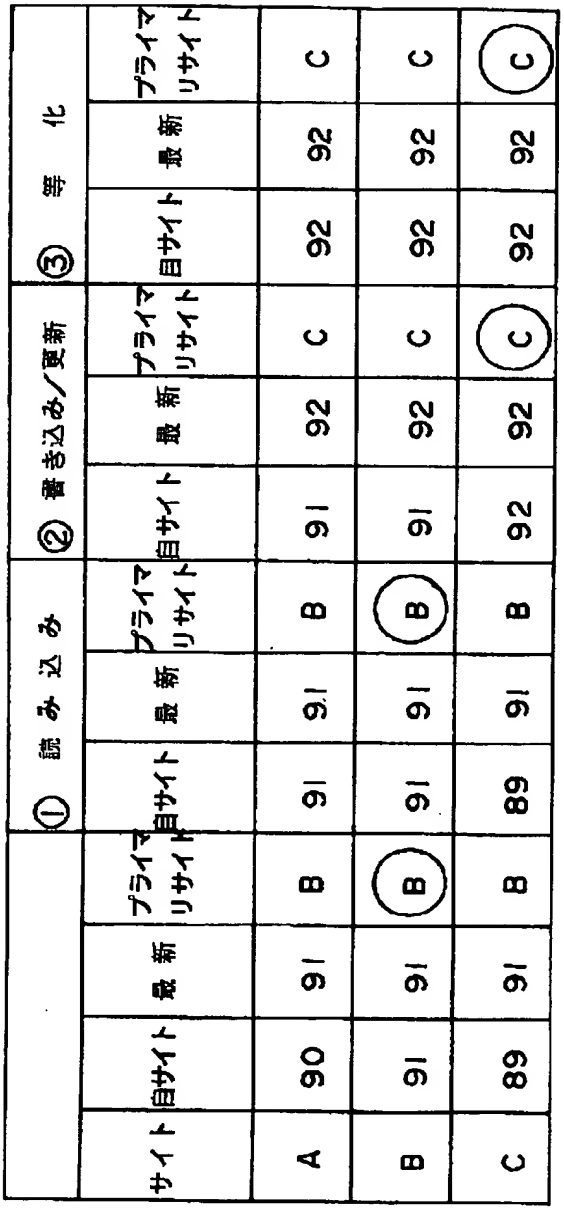
【図７】


ロック
トランザクション
A
プライマリコピー
ロック
ロック
コピー
コピー
トランザクション
C
トランザクション
B

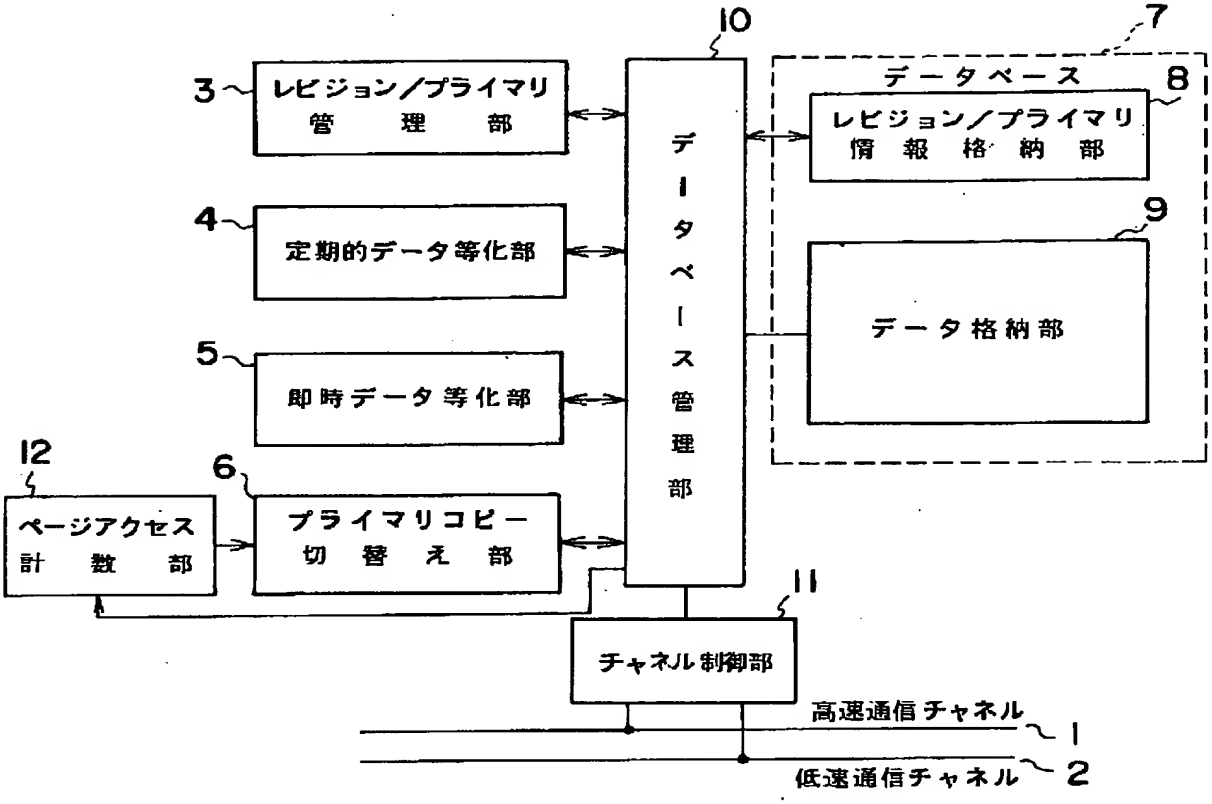
【図３】


サイト A
サイト B
7
c
90 91 B
古い
最新
90 91 Ⓑ
プライマリコピー
9
8
読み
502
端末装置
サーバー
503
①
e
1
2
②
③
a
b
d
f
書き／更新
サイト C
89 91 B

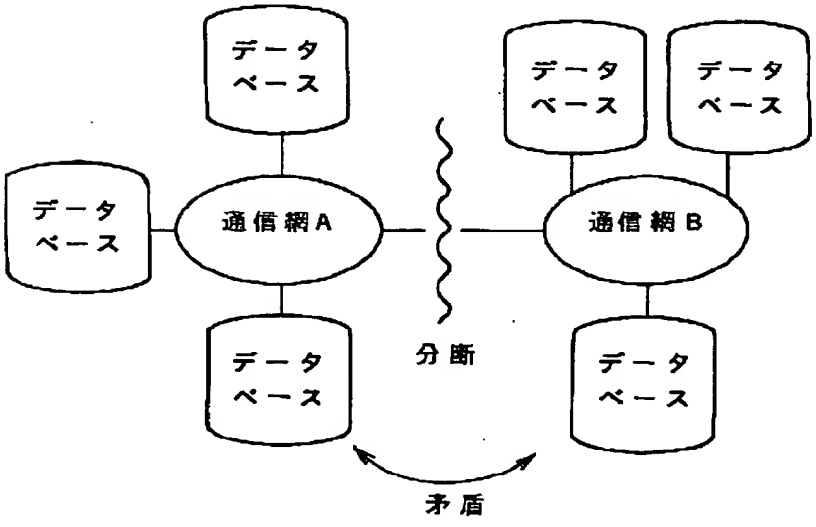
【図４】


サイト
自サイト
最新
プライマリサイト
①読み込み
自サイト
最新
プライマリサイト
②書き込み／更新
自サイト
最新
プライマリサイト
③等化
自サイト
最新
プライマリサイト
A
90
91
B
91
91
B
92
C
92
C
B
91
91
B
91
91
B
91
92
C
92
C
C
89
91
B
89
91
B
92
92
C
92
C

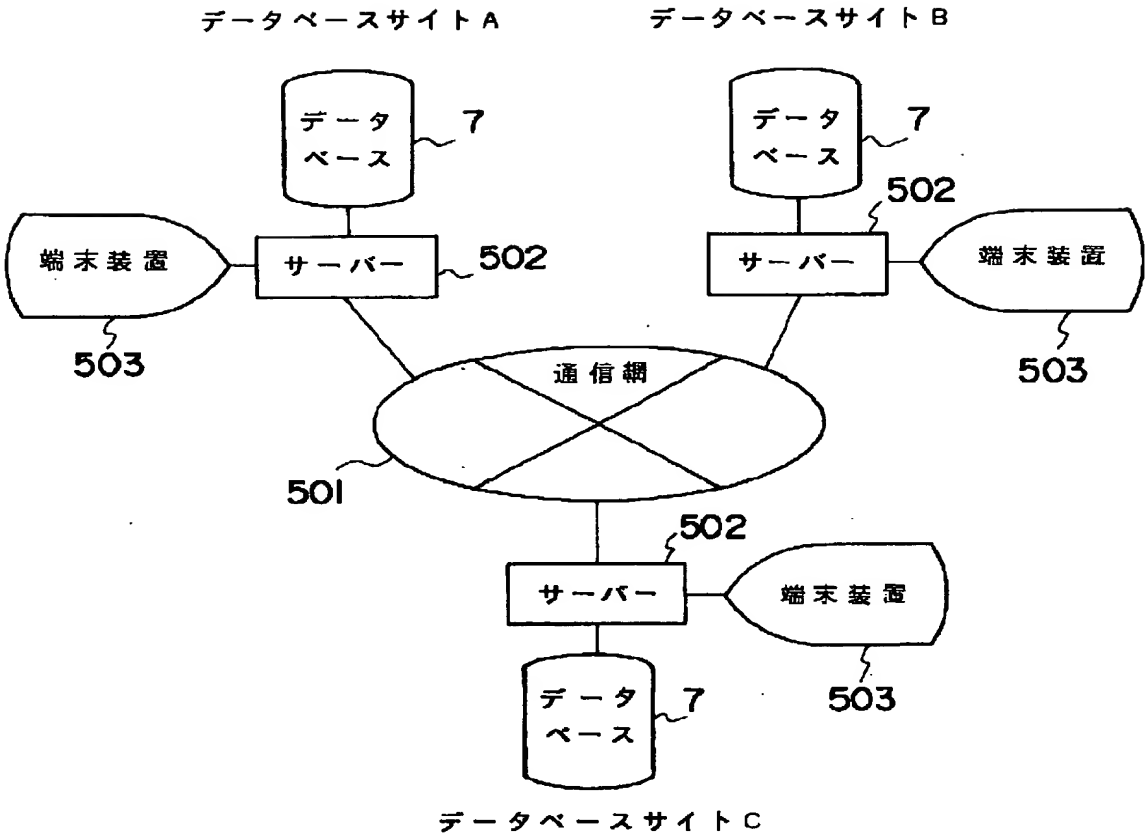
【図５】


10
7
データベース
8
3
レビジョン／プライマリ
管理部
レビジョン／プライマリ
情報格納部
4
定期的データ等化部
9
データ格納部
5
即時データ等化部
データベース管理部
12
6
ページアクセス
計数部
プライマリコピー
切替え部
11
チャネル制御部
高速通信チャネル
1
低速通信チャネル
2

【図８】


データベース
データベース
データベース
データベース
通信網A
通信網B
データベース
データベース
分断
矛盾

【図６】


データベースサイトＡ
データベースサイトＢ
データベース
7
端末装置
サーバー
502
503
通信網
501
データベースサイトＣ